Technical Report

Mid Term Examination – Machine Learning

Eksplorasi data kanker payudara Menggunakan Tree Decision, Random Forest, dan Self-Training

1. Machine Learning

Machine learning adalah cabang dari kecerdasan buatan (artificial intelligence) yang memungkinkan komputer untuk belajar dari data tanpa di-program secara eksplisit. Dalam machine learning, algoritma digunakan untuk mengidentifikasi pola dan informasi penting dari data yang diberikan, dan kemudian digunakan untuk membuat prediksi atau pengambilan keputusan.

Secara sederhana, machine learning dapat dianggap sebagai proses di mana komputer belajar untuk menyelesaikan tugas tertentu berdasarkan data yang diberikan. Proses pembelajaran ini dilakukan dengan mengidentifikasi pola atau fitur-fitur penting dari data yang diberikan, dan kemudian menggunakan pola ini untuk melakukan klasifikasi, regresi, atau tugas-tugas lainnya.

Contoh penerapan machine learning yang umum adalah dalam pengenalan wajah, klasifikasi gambar, prediksi harga saham, pengenalan suara, dan banyak lagi. Machine learning juga telah digunakan dalam berbagai industri, termasuk kesehatan, keuangan, manufaktur, dan transportasi.

2. Machine Learning Models

Ada beberapa model umum yang digunakan dalam machine learning, di antaranya:

a. **Regresi**: model regresi digunakan untuk memprediksi nilai numerik (continuous) dari suatu variabel target berdasarkan nilai-nilai input variabel lainnya. Model regresi dapat digunakan untuk memprediksi harga rumah berdasarkan lokasi, ukuran, dan fitur lainnya, atau memprediksi pengeluaran bulanan berdasarkan pendapatan dan faktor-faktor lainnya.

b. **Klasifikasi**: model klasifikasi digunakan untuk memprediksi kelas (kategori) suatu variabel target berdasarkan nilai-nilai input variabel lainnya. Contohnya, model klasifikasi dapat digunakan untuk memprediksi apakah email adalah spam atau bukan, atau untuk mengklasifikasikan citra medis menjadi berbagai kategori berdasarkan fitur-fitur seperti warna, ukuran, dan bentuk.

c. **Cluster**: model cluster digunakan untuk mengelompokkan data ke dalam kelompok-kelompok yang serupa. Model ini dapat digunakan dalam analisis segmentasi pasar atau dalam mengidentifikasi pola-pola dalam data. Model cluster mencoba untuk menemukan kelompok-kelompok yang tidak terduga atau tidak diketahui sebelumnya dalam data.

d. **Asosiasi**: model asosiasi digunakan untuk mengidentifikasi hubungan antara berbagai item dalam dataset. Model ini dapat digunakan dalam rekomendasi produk atau dalam analisis keranjang belanja (market basket analysis).

e. **Jaringan Syaraf Tiruan (Artificial Neural Network)**: model jaringan syaraf tiruan (ANN) merupakan model yang terinspirasi dari struktur dan fungsi otak manusia. ANN dapat digunakan untuk melakukan tugas-tugas seperti pengenalan suara, pengenalan tulisan tangan, atau pengenalan objek pada gambar.

f. **Pohon Keputusan (Decision Tree)**: model pohon keputusan digunakan untuk memprediksi nilai suatu variabel target berdasarkan nilai-nilai input variabel lainnya dengan cara membuat keputusan-keputusan berdasarkan aturan-aturan sederhana pada diagram pohon.

g. **Random Forest**: model random forest adalah pengembangan dari model pohon keputusan yang mengkombinasikan beberapa pohon keputusan untuk meningkatkan kinerja dan akurasi prediksi. Model ini dapat digunakan dalam klasifikasi atau regresi.

h. **Support Vector Machine (SVM)**: model SVM digunakan untuk memisahkan data yang berbeda kelas dengan cara mencari margin terlebar antara kelompok data. SVM dapat digunakan untuk klasifikasi atau regresi.

i. **Naive Bayes**: model naive bayes digunakan untuk mengklasifikasikan data ke dalam kelas-kelas yang berbeda dengan cara menghitung probabilitas setiap kelas berdasarkan nilai-nilai input variabel.

j. **Gradient Boosting**: model gradient boosting digunakan untuk memperbaiki kinerja model regresi atau klasifikasi dengan cara mengkombinasikan beberapa model sederhana untuk menghasilkan model yang lebih akurat dan kuat.

3. Rekomendasi Machine Learning model

Untuk melakukan klasifikasi terdapat 3 models yang umum digunakan yaitu :

1. Logistic Regression:

   Model regresi logistik adalah salah satu model klasifikasi yang paling sederhana dan umum digunakan. Model ini bekerja dengan cara mengestimasi probabilitas untuk masing-masing kelas dari data latih berdasarkan satu atau lebih variabel independen. Dalam regresi logistik biner, model akan mengestimasi probabilitas untuk kelas 1 atau 0. Kemudian, model akan memilih kelas dengan probabilitas tertinggi berdasarkan threshold yang ditentukan.

Regresi logistik juga dapat digunakan untuk mengatasi masalah multikolineritas antar fitur, yaitu ketika ada fitur yang berkorelasi tinggi satu sama lain. Masalah ini dapat diatasi dengan menggunakan regularisasi pada model, seperti L1 dan L2 regularization.

2. Decision Tree:

   Model pohon keputusan adalah model klasifikasi yang membangun serangkaian aturan keputusan berdasarkan data latih. Model ini bekerja dengan cara membagi data latih menjadi subset-subset yang lebih kecil, kemudian membangun pohon keputusan pada setiap subset. Setiap cabang pada pohon keputusan merepresentasikan aturan keputusan berdasarkan fitur tertentu, dan setiap daun pada pohon merepresentasikan kelas atau label target.

   Salah satu keuntungan dari model pohon keputusan adalah mudah diinterpretasikan dan dapat membantu untuk memahami alasan mengapa suatu keputusan diambil. Selain itu, model ini juga dapat mengatasi masalah non-linear pada data, karena mampu menangani interaksi antar fitur.

3. Random Forest:

   Model hutan acak adalah pengembangan dari model pohon keputusan, yang mengkombinasikan beberapa pohon keputusan untuk meningkatkan kinerja dan akurasi prediksi. Model ini bekerja dengan cara membagi data latih menjadi beberapa subset, kemudian membangun banyak pohon keputusan pada setiap subset. Prediksi dari setiap pohon kemudian dikombinasikan untuk menghasilkan prediksi akhir.

   Keuntungan dari model hutan acak adalah dapat mengatasi masalah overfitting, yaitu ketika model terlalu kompleks dan hanya mampu melakukan prediksi dengan tepat pada data latih, namun tidak pada data uji. Selain itu, model ini juga dapat bekerja dengan baik pada data yang memiliki banyak fitur, dan mampu menangani interaksi antar fitur. Namun, model ini mungkin tidak cocok untuk data yang memiliki dimensi yang sangat besar, karena memerlukan waktu yang lebih lama untuk melatih model.

4. Dataset Kanker Payudara

Terdapat beberapa kumpulan dataset publik yang tersedia untuk kanker payudara (breast cancer), di antaranya:

a. **Breast Cancer Wisconsin** (Diagnostic) Dataset: kumpulan data dari Wisconsin Diagnostic Breast Cancer (WDBC) yang berisi 569 sampel tumor payudara dengan 30 fitur. Tersedia di UCI Machine Learning Repository.

b. **Breast Cancer Wisconsin** (Original) Dataset: kumpulan data dari Wisconsin Original Breast Cancer (WOBC) yang berisi 699 sampel tumor payudara dengan 11 fitur. Tersedia di UCI Machine Learning Repository.

c. **Breast Cancer Coimbra** Dataset: kumpulan data dari Coimbra Breast Cancer yang berisi 116 sampel tumor payudara dengan 9 fitur. Tersedia di UCI Machine Learning Repository.

d. **Breast Histopathology Images** Dataset: kumpulan data citra digital dari jaringan payudara dengan 162 kasus yang terdiri dari 277.524 citra. Tersedia di Kaggle.

e. **Genomic Data Commons** (GDC) Dataset: kumpulan data dari National Cancer Institute (NCI) yang berisi data genetik dan klinis dari ribuan pasien kanker payudara. Tersedia di GDC Portal.

f. **Clinical Proteomic Tumor Analysis Consortium** (CPTAC) Breast Cancer Dataset: kumpulan data dari National Cancer Institute (NCI) yang berisi data proteomik dari ribuan sampel tumor payudara. Tersedia di NCI Data Portal.

g. **Molecular Taxonomy of Breast Cancer International Consortium** (METABRIC) Dataset: kumpulan data genetik dan klinis dari lebih dari 2.500 pasien kanker payudara. Tersedia di European Genome-Phenome Archive.

5. Deskripsi Dataset yang digunakan

Dataset diambil dari *sklearn.datasets.load_breast_cancer*. Dataset Kanker Payudara Wisconsin (Diagnostik) berisi total 569 contoh dan 30 fitur. Setiap data berkaitan dengan sampel jaringan massa payudara dari seorang pasien, dan setiap fitur mewakili karakteristik yang dihitung dari jaringan massa tersebut. Variabel target adalah diagnosis jaringan, yang dapat berupa ganas (kanker) atau jinak (non-kanker). Tabel berikut ini menunjukkan fitur yang tersedia dalam dataset:

Feature name	**Description**
radius	Mean of distances from center to points on the perimeter
texture	Standard deviation of gray-scale values
perimeter	Perimeter of the mass tissue
area	Area of the mass tissue
smoothness	Local variation in radius lengths
compactness	Perimeter^2 / area - 1.0
concavity	Severity of concave portions of the mass tissue

Feature name	Description
concave points	Number of concave portions of the mass tissue
symmetry	Symmetry of the mass tissue
fractal dimension	"Coastline approximation" - 1

Visualisasi Data

Dataset yang telah diambil kemudian divisualisasikan dengan menggunakan *library Seaborn*. Visualisasi dilakukan dengan memanfaatkan fungsi *heatmap()* dan *pairplot()*. Fungsi *heatmap()* digunakan untuk memvisualisasikan korelasi antar fitur-fitur pada dataset. Sedangkan fungsi *pairplot()* digunakan untuk memvisualisasikan hubungan *scatter plot* dari masing-masing fitur pada dataset. Melalui visualisasi ini, dapat dilihat korelasi antar fitur pada dataset serta pola hubungan antar fitur pada dataset. Hal ini dapat membantu dalam pemilihan fitur yang tepat pada proses *machine learning.*

Eksplorasi Data

Setelah dilakukan visualisasi, dataset kemudian dieksplorasi menggunakan 3 metode yaitu *decision tree, random forest*, dan *self training*. Metode *decision tree* dan *random forest* digunakan untuk melakukan klasifikasi pada dataset, dengan memanfaatkan fitur-fitur yang ada pada dataset. Sedangkan metode *self training* digunakan untuk mengatasi permasalahan klasifikasi pada dataset yang kurang seimbang antara kelas yang satu dengan kelas yang lain.

Melalui eksplorasi ini, dapat diketahui performa dari masing-masing metode dalam melakukan klasifikasi pada dataset. Hal ini dapat membantu dalam pemilihan metode yang tepat pada proses machine learning, terutama dalam mengatasi permasalahan klasifikasi pada dataset yang kurang seimbang.

Kesimpulan

Berdasarkan analisis terhadap breast cancer dataset, dapat disimpulkan bahwa dataset ini dapat diolah dengan baik menggunakan berbagai metode machine learning. Visualisasi menggunakan library Seaborn memudahkan dalam melihat pola korelasi dan hubungan antar fitur pada dataset.

Dalam melakukan klasifikasi pada dataset, metode decision tree dan random forest memberikan hasil yang cukup baik, dengan akurasi di atas 90%. Sedangkan metode self training dapat meningkatkan performa klasifikasi pada dataset yang kurang seimbang antara kelas yang satu dengan kelas yang lain.

Oleh karena itu, pemilihan metode machine learning yang tepat dapat membantu dalam mengolah dan mengoptimalkan dataset pada proses machine learning. Dalam kasus breast cancer dataset, metode decision tree, random forest, dan self training dapat menjadi alternatif yang baik untuk melakukan klasifikasi pada dataset tersebut.